DOCUMENT RESUME

ED 473 808 TM 034 785

AUTHOR	Min, Kyung-Seok; Frank, Kenneth A.
TITLE	The Impact of Nonignorable Missing Data on the Inference of Regression Coefficients.
PUB DATE	2002-10-00
NOTE	46p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Columbus, Ohio, October 16-19, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Data Collection; *Regression (Statistics); *Research Methodology; *Statistical Inference
IDENTIFIERS	*Missing Data

ABSTRACT

Various statistical methods have been available to deal with missing data problems, but the difficulty is that they are based on somewhat restrictive assumptions that missing patterns are known or can be modeled with auxiliary information. This paper treats the presence of missing cases from the viewpoint that generalization as a sample does not fully represent the target population. An index is developed to detect the impact of missing data on the inference of regression coefficients in terms of statistical test/significance. It is considered that the population consists of two separable subpopulations, one in which a linear relationship among variables of interest differs and one in which a sample from the populations under represents or over represents one of subpopulations. In order to derive the index of the impact of missing data, four hypothetical situations of simple regression are considered, and the expansion to a multivariate situation is provided. In addition, the features of this index are discussed in comparison with other statistical methods for missing data such as propensity scores, nonparametric models, and Fail-Safe N. (Contains 1 table, 7 figures, and 36 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 473 808

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

K.-S. Min

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Impact of Nonignorable Missing Data on the Inference of Regression Coefficients

Kyung-Seok Min

Kenneth A. Frank

Michigan State University

Paper presented at the annual meeting of the Mid-Western Educational Research Association

October 17, 2002

Columbus, Ohio

TM034785

BEST COPY AVAILABLE

ABSTRACT

In social and behavioral science, the data often includes missing cases by reason of response refusal, data editing, attrition and so on. Various statistical methods have been available to deal with missing data problems. However, the difficulty is that they are based on somewhat restrictive assumptions that missing patterns are known or can be modeled with auxiliary information. In this paper, the presence of missing cases is treated from the view of generalization as a sample does not fully represent the target population. An index is developed to detect the impact of missing data on the inference of regression coefficients in terms of the statistical test/significance. In particular, it is considered that the population consists of two separable subpopulations, where a linear relationship among variables of interest differs, and a sample from the population under- or over-represents one of subpopulations. In order to derive the index of the impact of missing data, four hypothetical situations of simple regression are considered and its expansion to a multivariate situation is provided. In addition, the features of this index are discussed compared with other statistical methods for missing data such as propensity scores, nonparametric models and Fail-Safe N.

Key words: nonignorable missing, statistical inference, regression model

The Impact of Nonignorable Missing Data on the Inference of Regression Coefficients

1. Introduction

Frequently researchers confront difficulties regarding missing cases/values while conducting quantitative analyses. In social and behavioral science, rather than being fully complete, the data often include missing cases for reasons such as refusal of response, editing out of inappropriate values, attrition and so on. In other words, the sample data is selected on some bases that are not completely known (Wainer, 1989), which is called selection bias. In order to deal with this problem caused by missing cases, and to obtain unbiased and efficient estimates, researchers have categorized several patterns and mechanisms of missingness and have developed related analytic strategies (see, Cohen and Cohen, 1983; Little and Rubin, 1987, 1989; Lohr, 1999).

From the view of generalization, the missing data problem can be understood as a case where the sample does not fully represent the target population. The strong tradition of sampling design points out that randomization (probability sampling) is the way to ensure the external validity of the study (Lohr, 1999). However, what if we have some missing cases even under an attempted random sampling design such that selection bias exists? When the response tendency is related to unmeasured values, that is, there is systematic loss of observations, we still have factors that might distort statements of causal links and decrease the power of statistical inference (Birnbaum and Mellers, 1989; Greenland, 2000). Moreover, considering the fact that social and behavioral science usually relies on observational studies in which random sampling frameworks are quite rare, the situation is worse. That is, due to various causes of missing data, the obtained sample may represent only part of the targeted population and the other part of the

population is underrepresented (or not represented) unless there is a proper consideration for missing cases. This is the reason we emphasize high response rates as well as an appropriate sampling design in research practice.

Various statistical techniques have been developed to deal with missing data in the regression model. Most of them are conducted under the assumption that data are missing at random (MAR) or missing completely at random (MCAR). Therefore more attention needs to be paid to statistical results under the possibility that missingness is systematic rather than at random. Further, the collection of work on missing data has focused on the situation in which one of the predictor(s) or an outcome variable is missing, but not both of them which is the case dealt with here (see, Allison, 2000; D'Agostino and Rubin, 2000; Daniels and Hogan, 2000; Dehejia and Wahba, 1999; Little, 1992).

In this paper the focus lies on detecting the impact of unobserved/missing data on inference about regression coefficients in terms of statistical tests/significance. In particular, it is considered that the intended sample consists of two separable subgroups, across which a linear relationship among variables of interest (e.g., a regression coefficient of Y on X) differs, and an observed/obtained sample from the initial sampling framework under/over- represents one of the subgroups. In other words, unobserved cases belong to the underrepresented subgroup and they can be treated as missing data which can improve coverage/representativeness of the observed sample. Because of differences in two subgroups in terms of a linear relationship, these unobserved cases could change the inference about the relationship between Y and X if they were in the sample. Then it can be asked how many unobserved cases with certain statistical characteristics (which will be discussed later) would be needed to alter the initial statistical inference regarding a regression coefficient. This question will be answered by developing a

simple index that quantifies the impact of missing data as a ratio of two separable sample sizes for observed cases and unobserved cases. Therefore this index informs the robustness of the statistical inference about the regression coefficient.

In sections 2 and 3, we will outline the characteristics of missingness and the leading approaches to dealing with missing data problems. In section 4, we evaluate the behavior of regression coefficients with unobserved cases to quantify changes in the statistical inference under specific conditions. Also further expansion to a multiple regression model is presented in section 5 and an example is provided in section 6. In the final section, we discuss the similarity and difference of the proposed index with other statistical methods, and comment on the limitations of this study.

2. Missing Data Mechanisms: Ignorable vs. Nonignorable

Little and Rubin (1987) categorized missing data mechanisms into ignorable and non-ignorable. The key issue is whether or not missingness depends on the missing values. For the ignorable case, there are two types, missing completely at random (MCAR) and missing at random (MAR). MCAR means that missing data is not only independent of other variable(s) in a data set but also independent of the unobserved values. Therefore, the complete cases can be treated as a random subsample of the intended sample. MAR implies that missing does not depend on the unobserved values but is related to other variables. Since missing cases can be deleted or fixed by employing auxiliary information from other variable(s), missing data in both cases do not much distort statistical inferences, therefore they can be ignored.

On the other hand, a nonignorable missing case occurs when missingness somehow depends on unobserved values. As a famous example, people with very high or low income tend

to refuse reporting their actual amounts of income so that researchers may expect that the missing responses would be located at both far ends of an income distribution rather than equally distributed across the full range of incomes. Therefore they can't just ignore missing cases without further consideration about patterns of missing data.[1]

3. Statistical Methods for Missing Data

We now review three types of statistical methods for missing data with either ignorable or nonignorable missingness assumed.

Complete-case analysis and available-case analysis

The standard treatment of missing data in statistical packages is the complete-case analysis in which missing cases are simply discarded and traditional statistical methods are then conducted. It is also known as listwise deletion, and works well under the assumption of MCAR but may fairly reduce sample size even when missing is sparse across variables.

The available-case analysis includes all observed cases to estimate each individual parameter. It is known as pairwise deletion. For example, bivariate correlation coefficients are obtained from all available cases in each pair of variables. Although the available-case method is appealing in that maximal information is used, the estimated covariance matrix is not necessarily positive definite. This is of concern especially when independent variables are highly correlated in a regression model. Also, the fact that the sample size varies from parameter to parameter is another disadvantage for further analysis. Indeed, it makes the degrees of freedom for statistical inference ambiguous.

Simple Imputation

In simple imputation, missing values are filled in with an unconditional/conditional sample mean, and the resultant complete data set is analyzed in a general way. Unconditional mean imputation is a simple approach to impute missing values with observed sample mean while conditional mean imputation method obtains information for missing values from other observed variables (i.e., conditional mean given other related variables). In order to compensate for the uncertainty in imputing missing cases, the weighting method (e.g., weights proportional to the inverse of the response rates and the selection rates) can be adopted in calculation (Little and Rubin, 1987). However, weighting adjustments are usually used for each subject but not for each observation within a subject (Lohr, 1999).

Under the assumption of MAR or MCAR, these two methods are reported to generate unbiased estimates of regression coefficients but they do not account for uncertainty of imputing values, called imputation errors. It means that these two methods result in underestimated variances, reducing the standard errors of regression coefficients and then overstating the precision of the estimates.

Model-based Methods

Model based methods include the maximum likelihood method, the Bayesian approach and multiple imputation. In the maximum likelihood method, under the common assumption of a multivariate normal distribution with a mean vector $\vec{\mu}$ and a covariance matrix Σ, the factored likelihood method [2] is adopted to obtain parameters of a joint distribution (Gourieroux and Montfort, 1981). The Bayesian approach is to multiply the likelihood for the data by a prior distribution and the inference is based on the resultant posterior distribution. It is effective for

small or moderate sample size inference compared to the maximum likelihood method which requires fairly large sample sizes. The Bayesian approach, however, has been applied to multivariate problems with missing dependent variables but applications to missing predictors are limited (Guttman and Menzefricke, 1983; Little, 1992).

Rubin (1987) proposed multiple imputation (stochastic regression imputation, Little and Rubin, 1989) as a solution to the problem of underestimated variance (overstated precision) from simple imputation methods. Multiple imputation randomly draws more than 2 values from the conditional distribution to fill in missing values and then these multiple filled-in data sets are analyzed. So the estimated variance can reflect uncertainty in the imputation process by including two sources of variances: the average variance within each imputed data set and the variance between imputations.

It should be noted that first two methods above have been built under the condition that missing is completely at random (MCAR) or information that is needed to fill in missing values is obtained from other observed (MAR). Further, model-based methods are also restrictive in that assumed models should be correct in some sense (Allison, 2000; Wainer, 1989). Since the correct specification of missing data mechanisms (selection bias) is not easy or might be impossible to establish, sensitivity analysis for the specified models is commonly conducted along with model-based methods. In fact, sensitivity analysis is not popular in practice because of its computational intensity and various possibilities of interpretation of the results (Frank, 2000).

Another note is that the three methods have been applied to situations in which at least some variables are observed for each subject in the sample. As such these methods improve

estimation by utilizing on available data. In contrast, the approach here will be to inform inference relative to hypothetical cases that are missing on all variables for some subjects. The idea here is that if all observed subjects had responded to survey questions, it would change the statistical inference obtained from the observed sample. So concerns are about the coverage of an observed sample (Cohen and Cohen, 1983, 276-277) and then about generalization through statistical inferences.

The following section presents procedures to index the impact of completely missing cases on statistical inference in four hypothetical situations.

4. The Impact of Missing Data on the Inference about Regression Coefficients

When one wants to predict an outcome Y, based on a predictive variable X, a regression model is commonly employed to quantify a linear relationship. Unfortunately, one does not always have complete observations of all cases due to nonresponse, attrition, data editing, and so on. Sometimes both X and Y are not observed. When the initially intended sample consists of two separable groups (e.g., male and female, low SES and high SES, etc.) that are suspected to have different relationships between Y and X, it might be possible that the observed cases over-represent one group due to improper sampling or nonresponse. For example, one group is not included in the sampling framework or most subjects of one group refuse to respond. In this case critics might suspect inferences would be different between incompletely observed data and fully observed data (combined data with both observed and unobserved/missing cases).

To address this concern we will explore hypothetical circumstances under which inferences would be altered if cases of which all variables are unobserved were included in the sample. Suppose two separable subgroups have different relationships among variables of

interest; the regression coefficient for one group (which the observed sample belongs to) is large enough to be statistically significant but not for the other group (which unobserved sample belongs to). To deal with the concern about different linear relationships across subsamples, we need to detect differences in regression coefficients according to the degree to which an initial sample is observed.

We can start with the two following simple regression models,

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad \text{for observed cases,} \tag{1}$$

$$y_i = \beta_0^* + \beta_1^* x_i + e_i^*, \text{ for observed and unobserved cases,} \tag{2}$$

where y_i and x_i are values for subject i.

When the relationship between X and Y expressed by $\hat{\beta}_1$, is statistically significant in equation 1, critics may ask about $\hat{\beta}_1^*$ in equation 2 and doubt the validity of the statistical inference from equation 1. In order to compare the essential conditions differentiating the two models, three conditions will be assumed. In particular, means and variances of X and Y respectively are assumed to be the same for the observed and unobserved cases. Therefore, the difference between $\hat{\beta}_1$ and $\hat{\beta}_1^*$ is determined only by differences in two sample covariances. These two assumptions are merely typical assumptions of regression. Differences in means should be accounted for with covariate(s) and homogeneous variances are assumed for inference.

In addition, it is assumed that the covariance between X and Y in unobserved cases is zero. It should be noted that the value of zero is for the sample but not for the population. The value of zero is considered as a neutral location regarding variously possible values of the covariance for unobserved cases.

Based on the framework of two separable subsamples and zero covariance for the

unobserved sample, we can consider the following four hypothetical situations to which critics may react, according to the original statistical inference and whether data are added or removed from the originally observed sample.

Case 1. Adding cases with a null relationship between X and Y

Suppose we find a statistically significant linear relationship between X and Y, based on a regression coefficient $\hat{\beta}_1$, from the observed data set but the observed cases do not fully cover the initially intended sample. As is previously mentioned, it is considered that the initially intended sample consists of two separable subgroups, across which a linear relationship among variables of interest differs. An observed sample includes only part of the intended sample which has a strong relationship between X and Y, and unobserved cases that belong to the underrepresented part of the initial sample can be treated as missing data. Because of differences of two subgroups, these unobserved cases could change the inference about the relationship between Y and X if they were in the sample. Then the question is how many unobserved cases with a zero covariance between X and Y need to be added to the observed sample to change the statistical inference for β_1.

For example, there are three populations in Third International Mathematics and Science Study (TIMSS) which has investigated the relationships between various schooling factors and students' achievement. One internationally desired population for final year of secondary school is defined as *"all students in the final year of secondary school, with those having taken advanced mathematics courses and those having taken physics courses as two overlapping sub-populations"* (Dumais, 1998, p. 15). In order to make observed cases cover the initially intended sample and fully represent the target population, weighting methods are adopted in TIMSS. As a

result, subjects (e.g., students, classrooms, or schools) in sampling strata with lower response rates get more weight and subjects with higher response rates get less weight. Indeed, this weighting scheme holds only if nonresponses occur at random.[3] Without using weights we may ask whether the statistical inference about the relationships between various schooling factors and students' achievement from the observed sample would be altered if unobserved subjects were included/observed.

Case 2. Replacing part of the observed sample with cases with a null relationship between X and Y

From the same situation as case 1, assume we want to maintain the sample size because the sample size is directly related to the significant test (i.e., sampling error). So we need to replace some observed cases with unobserved null cases rather than to add null cases in order to improve sample coverage. Here, the question is how many cases with the originally significant relationship between X and Y need to be replaced with a null relationship to alter the statistical inference.

Back to TIMSS, the study of final year of secondary school targets students who are in the last grade of the secondary school system. If one wants to know about all school-leaving age group both in and out of schools, samples in TIMSS are not appropriate because people outside of school are not considered. To make the inference about the general population of the school-leaving age group rather than the school population in the final school year, one additionally needs to sample from a specific age group (e.g., 18 -19 years olds) who are not enrolled in school. We may expect higher response rates from in-school samples than from out-school samples because the school system is better for sampling and testing than private organizations

or groups of individuals without any common affiliation. In order to balance different response rates or maintain the original sample size (or sampling error rate), one may consider replacing part of the school sample with the out-school sample rather than combine them intact.[4]

Case 3. Removing cases with a null relationship between X and Y

In this case, we have a statistically nonsignificant inference for β_1 and the observed sample consists of two subsamples, one of which has a covariance of zero. Then the question is how many cases with a null relationship between X and Y need to be removed from the sample to change the statistical inference for β_1 from being nonsignificant to significant.

Case 4. Replacing null observed cases with cases of a significant relationship between X and Y

Starting with the same situation as case 3 with an initially nonsignificant relation between X and Y, we again want to maintain the sample size as in case 2. So we need to replace some observed null cases with cases of a nonzero relationship of X and Y. Here the question is how many cases with a significant relationship between X and Y are replaced to change the inference.

Crossing dichotomous statistical decisions for the observed sample and the consistency of sample sizes, these hypothetical situations are tabulated in Figure 1. The data structure of each case is provided in Figure 2. For all four cases, we assume that the intended sample is composed of two separable groups across which the relationships between X and Y are different in terms of the sample covariance. The different relationships are defined by whether or not the linear relationship is large enough to make the regression coefficient be statistically significant.

From the four hypothetical scenarios above, unobserved cases can be treated as being

missing and they are interpreted in two ways; 1) subjects who are selected as sample elements but do not respond or 2) subgroups in the population are inaccurately represented in the originally intended sample. For instance, case 1 can be understood as a situation of lack of responses such that an observed sample is biased toward one of the subgroups when the nonresponses have different characteristics from the responses. Case 2 can be treated as a problem of originally misspecified sample such that a researcher tries to meet the structure of the population with a given sample size or sampling error rate. Cases 3 and 4 correspond to cases 1 and 2, respectively, except that cases 3 and 4 start from the situations that the observed samples overrepresent the null group and therefore the original inferences are to not reject the null hypothesis.

If a relatively large number of missing cases is needed to alter the inference for β_1, it may be argued that the inference from the observed data is not sensitive to the sampling scheme and unobserved cases, and one can rely on the initial result. In other words, we can evaluate the robustness of statistical results from the observed data by the ratio of the sample size of observed cases to the sample size of the combined data with observed and unobserved cases that are needed to alter our statistical inference. This ratio index will be detailed and derived for each hypothetical example.

Case 1. Adding cases with null relationship between X and Y

In the previous simple regression model for the combined sample with both observed and unobserved cases, equation 2, the focus is on the estimate of β_1^* and its statistical test/significance compared with that of β_1 in equation 1. As is well known, the estimate of β_1^* is

the ratio of the covariance of X and Y to the variance of X. Since we have two subsets of data, observed and unobserved cases, the ratio of a regression coefficient estimate can be decomposed according to the data structure,

$$\hat{\beta}_1^* = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{\sum_{i=1}^{n+k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n+k}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=n+1}^{n+k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=n+1}^{n+k}(x_i - \bar{x})^2}$$

$$= \frac{SS_{xy(Observed)} + SS_{xy(Unobserved)}}{SS_{x(Observed)} + SS_{x(Unobserved)}}.$$

Let n and k be the observed sample size and the unobserved sample size, respectively. Under the assumed conditions (constant means and variances), the covariance of the observed data $\hat{\sigma}_{xy}$, is large enough to make $\hat{\beta}_1$ in equation 1 be statistically significant while the isolated covariance of missing data, σ^*_{xy}, is zero. Then the previous equation becomes

$$\hat{\beta}_1^* = \frac{n\hat{\sigma}_{xy} + k\sigma^*_{xy}}{(n+k)\hat{\sigma}_x^2} = \frac{n\hat{\sigma}_{xy}}{(n+k)\hat{\sigma}_x^2} \quad . \tag{3}$$

Note two initial subsample sizes, n and k, are used for the formula instead of $n-1$ and $k-1$ to make expressions simple. In order to do the statistical hypothesis test, we now need the standard error of estimate, SE ($\hat{\beta}_1^*$),

$$SE(\hat{\beta}_1^*) = \frac{\sqrt{\left(\frac{1}{n+k-2}\right)\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \sum_{i=n+1}^{n+k}(\hat{y}_i - y_i)^2\right)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=n+1}^{n+k}(x_i - \bar{x})^2}}$$

$$= \frac{\sqrt{\left(\frac{1}{n+k-2}\right)\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \sum_{i=n+1}^{n+k}(\hat{y}_i - y_i)^2\right)}}{\sqrt{(n+k)\hat{\sigma}_x^2}},$$

where $\hat{y}_i$ is a predicted value for the *i*th subject in equation 2 and $\hat{\sigma}_x^2$ is assumed constant across observed and unobserved samples.

The numerator inside the radical is

$$\left(\frac{1}{n+k-2}\right)\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \sum_{i=n+1}^{n+k}(\hat{y}_i - y_i)^2\right)$$

$$= \frac{\sum_{i=1}^{n+k}(\hat{y}_i - y_i)^2}{n+k-2} = (1 - \hat{R}^2_{combined})\hat{\sigma}_y^2 = (1 - \frac{\hat{\sigma}^2_{xy,combined}}{\hat{\sigma}_x^2\hat{\sigma}_y^2})\hat{\sigma}_y^2 . \quad (4)$$

The covariance between the two variables for the combined sample is defined only by observed cases since the isolated covariance for the unobserved data, σ^*_{xy} is set to zero.

$$\hat{\sigma}_{xy,combined} = \frac{\sum_{i=1}^{n+k}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{n+k} = \frac{n\hat{\sigma}_{xy} + k\sigma^*_{xy}}{(n+k)} = \frac{n\hat{\sigma}_{xy}}{(n+k)} . \quad (5)$$

By substituting the result of equation 5 into equation 4, we obtain

$$\left(1 - \frac{\hat{\sigma}^2_{xy,combined}}{\hat{\sigma}_x^2\hat{\sigma}_y^2}\right)\hat{\sigma}_y^2 = \left(1 - \frac{\left(\frac{n\hat{\sigma}_{xy}}{(n+k)}\right)^2}{\hat{\sigma}_x^2\hat{\sigma}_y^2}\right)\hat{\sigma}_y^2 = \hat{\sigma}_y^2 - \left(\frac{\frac{n\hat{\sigma}_{xy}}{n+k}}{\hat{\sigma}_x}\right)^2 .$$

So the complete form of the standard error of the regression coefficient is,

$$SE(\hat{\beta}_1^*) = \frac{\sqrt{\hat{\sigma}_y^2 - \left(\frac{n\hat{\sigma}_{xy}}{\frac{n+k}{\hat{\sigma}_x}}\right)^2}}{\sqrt{(n+k)\hat{\sigma}_x^2}}. \qquad (6)$$

From equations 3 and 6, the test statistic of the regression coefficient for the combined data is

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\frac{n\hat{\sigma}_{xy}}{(n+k)\hat{\sigma}_x^2}}{\frac{\sqrt{\hat{\sigma}_y^2 - \left(\frac{n\hat{\sigma}_{xy}}{\frac{n+k}{\hat{\sigma}_x}}\right)^2}}{\sqrt{(n+k)\hat{\sigma}_x^2}}}.$$

This test statistic has a *t*-distribution with the degrees of freedom of $n+k-2$, under the condition that errors are independent, distributed normally and identically, and the null hypothesis is true.

To examine how large the supplemental data with null covariance must be to alter an inference, suppose the test statistic, $\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)}$, is equal to the critical value of *t*, which means that the regression coefficient for the combined data is just significant,

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\frac{n}{\sqrt{n+k}}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}}{\sqrt{\hat{\sigma}_y^2 - \left(\frac{n}{n+k}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}\right)^2}} = t_{critical} \ (df=n+k-2). \qquad (7)$$

In equation 7, $\hat{\sigma}_{xy}$, $\hat{\sigma}_y^2$, $\hat{\sigma}_x^2$ and n are known from the observed sample and the only undetermined term is the number of unobserved cases, k. The basic idea in equation 7 is to

calculate the number of unobserved cases with a null relationship, which brings the inference for the combined data to the level of being just significant. In other words, how many null supplemental cases would be needed to alter the initial statistical inference for the observed data?

To answer the question about the number of unobserved cases, k, we need to solve equation 7 with respect to the unobserved sample size, k. Let M represent the ratio of the combined sample size ($n+k$) to the observed sample size (n), which Frank (2001) called the **d**ilution **v**olume for **a**ugmented **d**ata (DVAD),

$$M = \frac{(n+k)}{n} > 1, \text{ whenever } k > 0.$$

By substituting M into equation 7, we get

$$\frac{\sqrt{Mn}\hat{\sigma}_{xy}}{\sqrt{M^2\hat{\sigma}_x^2\hat{\sigma}_y^2 - \hat{\sigma}_{xy}^2}} = \frac{r\sqrt{Mn}}{\sqrt{M^2 - r^2}} = t_{critical} ,$$

$$\Rightarrow t^2M^2 - nr^2M - t^2r^2 = 0 . \qquad (8)$$

When we solve equation 8 with respect to M, the index is a function of the correlation of X and Y ($r = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x\hat{\sigma}_y}$) of the observed data, the observed sample size (n), and the critical value of t ($t=t_{critical}$ with the degrees of freedom of $n+k-2$).

$$M = \frac{(n+k)}{n} = \frac{nr^2 \pm \sqrt{n^2r^4 + 4t^4r^2}}{2t^2} . \qquad (9)$$

Because M is always larger than 1, the negative solution for M is to be ignored.[5] This index indicates the ratio of sample sizes that makes the regression coefficient for the combined data be just significant. It should be noted that M is a scale-free measure because it is a ratio of two sample sizes. Therefore we can easily compare the values of M across different samples to

evaluate the robustness of the inference.

In addition, equation 9 is easily transformed into the solution for k, the number of supplemental cases with covariance of zero,

$$k = Mn - n$$
$$= \frac{n^2r^2 + n\sqrt{n^2r^4 + 4t^4r^2}}{2t^2} - n.$$

Figure 3 shows that how large unobserved data should be in order to change the inference with two different observed sample sizes (n=28, 84)[6] at .05 level, according to values of the correlation for the observed data. Note that the critical t-value in equation 9 is determined by the combined sample size ($n+k-2$). When the observed sample size is already large enough (e.g., larger than 100), the t distribution is close to the standard normal distribution that 1.96 is used as the critical value of t at the .05 level. When the observed sample size is less than 100, the critical t value for the combined data depends on both n and k. In order to make a graphic representation like Figure 3, we used critical t values with degrees of freedom of 26 and 82 for the observed sample sizes of 28 and 84, respectively. Therefore, for these two relatively small samples, the critical t values are somewhat conservative since the combined sample size should be larger than 28 and 84.[7] Figure 3 shows that the combined correlation coefficient (or the regression coefficient) becomes nonsignificant when M is located above each curve with a given sample size. Note that M is less than 1 when the original correlation is not significant; a correlation coefficient between -.21 and .21 for the observed sample size 84, and between -.36 and 36 for the observed sample size 28. These situations are not considered for case 1.

The evaluation of the index M should be based on knowledge about the population structure or data collecting procedures. For example, if unobserved/missing cases represent a half of the intended sample or the response rate of the sample is 50%, we may interpret a calculated

index with a given sample size and correlation coefficient against the line of 2. Here the criterion number 2 is obtained by the ratio of the initially intended sample size to the actually observed sample size. So when a value of the index is less than 2, we may say that the misspecified sample with response rate 50% may have altered the initial inference. In other words, if the number of supplemental cases with a null relationship which brings down the regression coefficient to the level of being just significant is relatively small the inference from the originally observed sample is not robust to the impact of missing data.

Figure 3 also indicates that the ratio of the combined sample size to the observed sample size increases as the observed correlation becomes greater. The absolute value of the slope of the tangent onto the curve is greater when the observed sample size, n, is larger. Therefore, it takes proportionally more cases with zero covariance to alter an inference the larger the observed correlation and/or the observed sample size.

Since the critical t value depends on the significant level (α level), the ratio index M, in equation 9 depends on the significance level such as .05 and .01. Figure 4 shows how the ratio varies with the observed sample size of 84 for two significance levels. As is expected, the more restrictive the significance level (the smaller α), the less null missing data is needed to alter the initial statistical inference. Also slopes onto curves are not the same for the two α's and the .05 level has steeper tangent lines than the .01 level. All these indicate that our statistical inference is more sensitive to unobserved cases as the significance level is more restrictive.

In summary, the ratio index of the combined sample size and the observed sample size in equation 9 becomes larger (i.e., more robust inference) as an observed sample size (n) is larger, the correlation coefficient (r) is larger, and the critical value of t is smaller.

Case 2. Replacing part of observed sample with cases of null relationship between X and Y

In order to avoid the effect of change in sample size as a result of adding supplemental cases with zero covariance, we now consider replacement of observed cases with null cases (refer to Figures 1 and 2).

After replacing k out of n observed cases, the test statistic for $\hat{\beta}_1^*$ in equation 2 is similar to equation 7 which is for the situation of adding unobserved null cases,

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\frac{(n-k)}{\sqrt{n}}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}}{\sqrt{\hat{\sigma}_y^2 - \left(\frac{(n-k)}{n}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}\right)^2}} = t_{critical} \ (df=n\text{-}2) .$$

Compared with equation 7, it is noted that n and $n+k$ are replaced by $n-k$ and n, respectively because we are replacing k null cases rather than adding them such that the original sample size, n is maintained before and after replacement.

To obtain the new index, let M be equal to $\frac{n-k}{n}$ which is smaller than 1 by definition.

Define,

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\sqrt{n}\, M\, \hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_x^2\hat{\sigma}_y^2 - M^2\hat{\sigma}_{xy}^2}} = t \ ,$$

then divide both sides of the second equal sign by $\hat{\sigma}_x\hat{\sigma}_y$,

$$\Rightarrow nr^2 M^2 = t^2(1 - r^2M^2) \quad \Rightarrow M^2 = \frac{t^2}{(n+t^2)r^2} ,$$

where r is the correlation of X and Y for the initially observed n cases.

Then,

$$M = \frac{n-k}{n} = \pm\sqrt{\frac{t^2}{(n+t^2)r^2}} \quad . \tag{10}$$

Since the index M is larger than 0 and smaller than 1 by definition, take the positive of the root as a unique solution for M. The meaning of M is not the same as the previous index of case 1. While the index M for case 1 is the ratio of the combined sample size to the initially observed sample size, the index for case 2 is the ratio of the preserved sample size to the initially observed sample size. In addition, a more meaningful ratio is easily obtained by a simple manipulation of equation 10 such as $k/n = 1 - M$, the proportion of the observed sample of n that must be replaced to alter the inference.

The relation of r and M in equation 10 is drawn in Figure 5. The three curves show that the regression coefficient is just significant after replacing k null cases and it becomes nonsignificant when the ratio (M) is located below these curves for each sample size. The part of three curves above the line of M equal to 1 occurs where the initial correlation of X and Y is not significant, therefore this is not considered here but it will be discussed in case 4; observed correlation between -.36 and .36 for the sample size of 28, between -.21 and .21 for the size of 84, and between -.07 and .07 for the size of 783.

As described in case 1, the evaluation of the index M should be based on information on the population structure or data collecting procedures. Again, if observed cases represent only a half of the initially intended sample or the response rate of the sample is 50%, we may use the line of .5 to evaluate the index obtained from the observed sample. When a value of the index is larger than .5, it indicates that the misspecified sample may have altered the original inference.

Case 3. Removing cases with a null relationship between X and Y

So far we have considered situations where the observed sample is obtained from one of subgroups then calculated the number of cases with null covariance needed to alter the inference. On the other hand, we may consider the situation in which the regression coefficient in equation 1 is not significant and the observed sample consists of two separable groups in terms of different relationships between X and Y. It can be asked again how many cases of a null relationship between X and Y should be removed to change the initial statistical inference when a subgroup with null relationship is overrepresented in the observed sample.

To address this problem, we need to slightly modify equation 7 and redefine symbols. In the observed sample of $n+k$, define n as the number of cases with a significant relationship between X and Y and k as the number of cases with null relationship (i.e. a covariance of 0 with constant means and variances), of which l cases need to be removed to alter the initially nonsignificant relationship. Therefore, the possible range of l is zero to k, and $n+k-l$ will be the resultant sample size of which the covariance is expected to be large enough to make the linear relationship between X and Y statistically significant. After part (l) of k cases with null relationship are removed, the test statistic for the regression coefficient in equation 2 can be expressed as below

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\frac{n}{\sqrt{n+k-l}}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}}{\sqrt{\hat{\sigma}_y^2 - \left(\frac{n}{n+k-l}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}\right)^2}} = t_{critical} \quad (df = n+k-l-2) \tag{11}$$

$$\Rightarrow n\left(\frac{n}{n+k-l}\right)^2 r_n^2 = t^2 - t^2\left(\frac{n}{n+k-l}\right)^2 r_n^2$$

$$\Rightarrow \left(\frac{n^2}{n+k-l}\right)\left(\frac{n+k}{n}\right)^2 r_{(n+k)}^2 = t^2 - \left(\frac{n}{n+k-l}\right)^2\left(\frac{n+k}{n}\right)^2 t^2 r_{(n+k)}^2 .$$

where $r_{(n+k)}$ is the initially observed correlation coefficient of sample size $n+k$, and r_n is the correlation coefficient only for the subgroup of size n in which the covariance is assumed to be large enough to make the linear relationship significant.

The relation between the initially observed correlation coefficient $r_{(n+k)}$ and that for a subgroup n in which the correlation is significant r_n, is determined as,

$$r_{(n+k)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) + \sum_{i=n+1}^{n+k}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n+k}(x_i-\bar{x})^2 \sum_{i=1}^{n+k}(y_i-\bar{y})^2}}$$

$$= \frac{n\sigma_{xy} + k\sigma^{*}_{xy}}{(n+k)\hat{\sigma}_x\hat{\sigma}_y} \qquad (\text{since } \sigma^{*}_{xy} = 0)$$

$$= \frac{n}{n+k}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x\hat{\sigma}_y} = \frac{n}{n+k} r_n$$

$$\Rightarrow r_n = \frac{n+k}{n} r_{(n+k)} \cdot$$

Here $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are common variances for subgroups n and k, and $\hat{\sigma}_{xy}$ is the covariance of sample n with a significant relationship. After replacing the observed correlation $r_{(n+k)}$ for r_n in the result of equation 11 and setting the index M equal to $\frac{n+k-l}{n+k}$, we obtain,

$$M = \frac{n+k-l}{n+k} = \frac{(n+k)r^2_{(n+k)} \pm \sqrt{(n+k)^2 r^4_{(n+k)} + 4t^4 r^2_{(n+k)}}}{2t^2} \cdot$$

This final formula is the same as equation 9 of case 1 except that the initial sample size is $n+k$ rather n, and the index, M should be smaller than 1. Also the meaning of M is the same as in case 1, the ratio of the resultant sample size and the original sample size. We again have some

uncertainty for the critical value of t of this formula because it depends on both known values (n and k) and unknown value (l). As in case 1, we can use conservative degrees of freedom for a resultant sample to get the index when an initial sample size is small.

Figure 6 shows the relationship between the index M and observed correlation coefficient. This figure is the eliminated part of Figure 3, where the index is less than 1 because the observed correlation is not statistically significant; correlation coefficients between -.21 and .21 for the observed sample size of 84, and between -.36 and 36 for the sample size 28. From Figure 6, we are informed that more null cases, relative to the observed sample size ($n+k$), need to be removed as the correlation and/or the observed sample become smaller.

Case 4. Replacing null observed cases with cases of a significant relationship between X and Y

Now consider replacing l cases of a null relation between X and Y in the initially observed data, for which $\hat{\beta}_1$ is statistically nonsignificant, as in case 3. The total sample size, $n+k$ does not change since part of the sample is replaced rather than removed. After replacing null cases, the test statistic of $\hat{\beta}_1^*$ is also similar to previous ones:

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{\frac{n+l}{\sqrt{n+k}}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x}}{\sqrt{\hat{\sigma}_y^2 - \left(\frac{n+l}{n+k}\frac{\hat{\sigma}_{xy(n-k+l)}}{\hat{\sigma}_x}\right)^2}} = t_{critical} \ (df \text{ of } n+k-2) \ ,$$

$$\Rightarrow (n+k)\left(\frac{n+l}{n+k}\right)^2 r_n^2 = t^2 - t^2\left(\frac{n+l}{n+k}\right)^2 r_n^2$$

$$\Rightarrow (n+k)\left(\frac{n+l}{n}\right)^2 r_{(n+k)}^2 = t^2 - \left(\frac{n+l}{n}\right)^2 t^2 r_n^2 , \qquad (\text{since } r_n = \frac{n+k}{n} r_{(n+k)}).$$

Let the index M equal $\frac{n+l}{n}$, then

$$M = \frac{n+l}{n} = \pm\sqrt{\frac{t^2}{(n+k+t^2)r_{(n+k)}^2}}$$

where $r_{(n+k)}$ and r_n are the initially observed correlation coefficient and that of a subgroup of n, respectively as before.

Use the positive solution for M and note that this solution is the same as in case 2, equation 10 except for the difference in initial sample sizes ($n+k$ vs. n). The meaning of the index M is the ratio of cases with a significant relationship after and before replacement. In addition, a more useful meaning is obtained by a simple manipulation such as $l/n = M-1$, the proportion of replacing cases among the observed significant cases, n.

The relationship between the index M and observed correlation in case 4 is represented in Figure 7. Again, this figure is the eliminated part of Figure 5, where the index is larger than 1; observed correlation coefficients between -.36 and .36 for the sample size of 28, between -.21 and .21 for the size of 84, and between -.07 and .07 for the size of 783.. Figure 7 informs that more null cases, relative to the observed sample size ($n+k$), need to be replaced as the correlation and/or the observed sample size become smaller.

5. Expansion to Multiple Regression Cases

Only simple regression models are considered in section 4 but the same procedure may apply to the multiple regression model that includes several covariates as,

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \cdots + \beta_{p+1} z_{pi} + e_i$, for observed cases,

$y_i = \beta_0^* + \beta_1^* x_i + \beta_2^* z_{1i} + \cdots + \beta_{p+1}^* z_{pi} + e_i^*$, for observed and unobserved cases.

These models include covariates (z's) but the relationship of X and Y (β_1 and β_1^*) is still of primary interest. Define **Z** as a vector of $z_1, z_2, \cdots, z_p$ then the significance test of $\hat{\beta}_1^*$ for the combined data is

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{r_{y(x.z)combined}}{\sqrt{\frac{1-R_{y.xz}^2}{n+k-(p+2)}}},$$

where, $r_{y(x.z)combined}$ $(= \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1-r_{xz}^2}})$ indicates the semi-partial correlation between Y and X for the combined data, in which the common variance between X and **Z** is removed, p is the number of covariates (**Z**), and $R_{y.xz}^2$ is the squared multiple correlation coefficient. In order to simplify the formula and derive the index M, the following manipulations are incorporated.

1) When the sample size $(n+k)$ is fairly large relative to the number of predictors (p) we may use $n+k$ as degrees of freedom instead of $n+k-(p+2)$ in the formula.

2) The overall squared multiple correlation coefficient $R_{y.xz}^2$, is decomposed into two terms like stepwise regression methods,

   $R_{y.xz}^2 = R_{yz}^2 + r_{y(x.z)}^2$.

3) When we assume that the covariance of X and Y is zero for unobserved cases after removing the common variance between X and **Z**, two semi-partial correlation coefficients for the combined data and the observed data have a relationship as below,

$$r_{y(x.z)combined} = \frac{n}{n+k} r_{y(x.z)observed}.$$

This functional relationship between two semi-partial correlation coefficients is similar to the relationship between two correlation coefficients in cases 3 and 4.

Then define,

$$\frac{\hat{\beta}_1^*}{SE(\hat{\beta}_1^*)} = \frac{r_{y(x.z)combined}}{\sqrt{\frac{1-R^2_{y.xz}}{n+k-(p+2)}}} \approx \frac{\frac{n}{n+k} r_{y(x.z)observed}}{\sqrt{\frac{1-(R^2_{yz} + \frac{n}{n+k} r^2_{y(x.z)observed})}{n+k}}} = t$$

$$\Rightarrow t^2(1-R^2_{yz})M^2 - nr^2_{y(x.z)observed} M - t^2 r^2_{y(x.z)observed} = 0. \qquad (12)$$

where M is the ratio of sample sizes, $\frac{n+k}{n}$ and $r_{y(x.z)observed}$ is a semi-partial correlation coefficient between X and Y for the observed cases.

While the joint distribution of X and Y determines the relationship between X and Y in a simple regression model, we have to deal with the conditional joint distribution of X and Y in a multiple regression model because the model includes covariates. Therefore, in order for equation 12 to hold, the previous three conditions (consistent means and variances and zero covariance for the missing data) are again assumed after covariates, **Z** are controlled for.

As a result, equation 12 is very similar to equation 7 in case 1 except that a semi-partial correlation between X and Y is used and a new term reflecting the relationship between Y and **Z** is added $(1-R^2_{yz})$. If there is no linear relationship between covariates, and X and Y ($R_{yz} = 0$ and $r_{y(x.z)} = r_{yx}$) then equations 7 and 12 are the same.

Solving equation 12 for M, we obtain

$$M = \frac{(n+k)}{n} = \frac{nr^2_{y(x.\mathbf{z})observed} \pm \sqrt{n^2 r^4_{y(x.\mathbf{z})observed} + 4(1-R^2_{y\mathbf{z}})t^4 r^2_{y(x.\mathbf{z})observed}}}{2t^2(1-R^2_{y\mathbf{z}})}, \qquad (13)$$

where t is the critical value with the degrees of freedom of $[n+k-(p+2)]$.

From equation 13, we can determine that more null cases are needed to alter the original inference for β_1 as observed the sample size, n, is larger, the semi-partial correlation between X and Y is bigger, the correlation between Y and **Z** is bigger, and the critical t-value is smaller. The effect of covariates makes sense in that the weaker relationship between X and **Z** implies that the partial correlation between X and Y is stronger such that more null cases are needed to neutralize this relationship. Effects of other factors are the same ways as in equation 9 for a simple regression model.

Here, we have presented the multivariate extension only for case 1 of the simple regression using a semi-partial correlation, but other cases are also easily extended with similar procedures.

6. Example

Featherman and Hauser (1976) investigated gender inequality in terms of educational attainment and socioeconomic achievement. They compared gender differences over a decade with census data and obtained the following regression line for men in 1973,

EDU = 11.99 + .041 FAOCC - .922 FARM - .282 SIBS,

where EDU is the educational attainment (year), FAOCC is father's occupation (Duncan's index of socioeconomic status), FARM indicates farm origin (dummy variable), and SIBS is the number of siblings. The proportion of variation of educational attainment explained by this model is .25. The sample size is 23,591 and correlation coefficients among variables are

provided in Table 1.

The sampled population of Featherman and Hauser's study is "Married Spouse-Present (MSP) men in 1973." This means that unmarried, divorced, or widowed men were not included in the data. Therefore, it is not clear whether the relation between educational attainment and background variables, obtained from MSP sample, holds for all working men in 1973. As an example of the application of the index *M*, we could question about the relationship between education attainment and father's occupation.[8] In order to determine the robustness of the regression coefficient of FAOCC for the overall population (working men in 1973), we may ask "How many cases must be added, in which there is a null relationship between EDU and FAOCC after FARM and SIBS are controlled for, to alter the inference?" To answer this question we can calculate the index *M* and interpret it in terms of population structures and sample representativeness.

Referring to equation 13, we can obtain the index *M* of 3591.87. It means that we need about 3591 times as many null missing cases as the size of original observations to change statistical inference for the regression coefficient of FAOCC. This large number of the index comes out because the initially large sample size (n=23,591) and the relatively strong relationship between EDU and FAOCC (r=.416) dominate equation 13.

The evaluation of this number, 3,591 should be based on information on the composition of the target population. Unless the MSP men represent only one 3,591th of the overall population and the non-MSP men have a very strong negative relationship between EDU and FAOCC, we can say that the positive and statistically significant relationship between EDU and FAOCC with FARM and SIBS controlled for, may hold for the overall population.[9]

In addition, this example indicates that one needs a very large number of supplemental

data to change the inference for regression coefficients obtained from an initially large sample. So it might be safe to say that the index *M* works better and gives more practical interpretation with small or moderate size samples.

7. Discussion

We have developed the index *M* which indicate the robustness of statistical inference for regression coefficients. This index is derived to quantify the question "How many unobserved cases with a null relationship are needed to alter the inference about the relationship between two variables in a regression model?" Further, we dealt with four hypothetical bivariate cases classified by the statistical significance and changes of sample sizes. It was also demonstrated that the index for simple regression cases is easily expanded to multivariate cases with similar processes and then we applied this index to an example in which the relationships between the educational attainment and three background variables are discussed.

In developing the index *M*, we have relied on the traditional hypothesis test procedure. A statistical decision of regression is dichotomous based on the critical value and the sampling distribution of the test statistic under the null hypothesis; we either reject or do not. This traditional hypothesis test has been criticized in that it results in a binary conclusion based on arbitrary cut points (.05 or .01), and it incorporates only the null hypothesis but not the alternative hypothesis (see, Hunter, 1997; Schmidt, 1996; Thompson, 1989). Even though this critique is persuasive in some points, the determination of policy frequently requires a binary decision at a certain cutoff point and this cutoff point can be discussed and reasonably determined (Frank, 2000).

The benefit of the index *M* is that it can provide the degree or quantity of the impact of

unobserved cases on the statistical inference of regression coefficients, even though it is based on an arbitrary significant level. The essence of evaluation of the impact is to comparing the index M with a certain criteria obtained from the knowledge about the population structure or data collecting procedures, and to decide whether the impact is probable or improbable.

Also the developed index is based on three assumptions (consistent means and variances, and zero covariance) that may not be often satisfied in practice. However, as we noted before, differences in means can be accounted for by including covariate(s) in a model , which adjusts different locations, and homogeneous variances is consistent with the standard assumptions of regression or analysis of variance (ANOVA). If there are heterogeneous variances, a sensitivity analysis across the possible range of variances must be employed rather than a simple index which is developed in this paper. However, the question is how one can define the range of an unobserved sample. For the third condition, it is also true that zero covariance of missing data is not a usual situation in practice. However, zero covariance can be treated as a neutral point among variously possible range of covariances of the unobserved data when we are not sure about their statistical features. This condition also provide a conservative criterion to evaluate the impact of the unobserved data when the covariance of unobserved cases is not zero and has the same sign as that of the observed data. As such index developed with zero covariance condition is based on the scenario that is in favor of critics' concerns that the unobserved data could change the inference obtained from the observed data. On the other hand, when the covariance of the unobserved data has the opposite sign of that of the observed data, the index M becomes a liberal measure and underestimates the impact of missing data. Further, due to the lack of knowledge about the unobserved cases, we may want to unconstrain the covariance of the unobserved data from zero. Then, the impact of the unobserved data on regression coefficients can be indexed as

a ratio of two covariances/correlations of observed and unobserved data but we need to control the sample size of unobserved data (Frank, 2001).

We index the impact of missing cases that are not observed in the regression model and this concern is conceptually related to the counterfactual statement on potential outcomes (see, Little and Rubin, 2000; Rosenbaum and Rubin, 1983b; Stone, 1993). The counterfactual argument originates from experimental designs, "What if each subject had been assigned in the treatment group and the control group and both outcomes were observed?" This subjunctive mood can be slightly modified in the missing data context as if a researcher had had valid responses from all sample elements or if the sample had been fully representative of all subgroups in the intended sample.

One can express this concern as following regression model with an interaction term,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_i + \beta_3 (I_i \times x_i) + e_i \ ,$$

where I_i is an indicator of whether the ith case is observed or not.

If $\hat{\beta}_3$ is meaningful or statistically significant, we might make incorrect inferences from the regression model as in equation 1.[10] Similar concerns about uncertainty of unobserved responses have appeared with different terms and statistical solutions such as propensity scores, nonparametric model and file drawer problems which we will discuss below.

Propensity Score

Propensity scores methods are about controlling confounding variables in order to consolidate causal inferences from the regression model or ANCOVA model (see, D'Agostino and Rubin, 2000; Dehejia and Wahba, 1999; Greenland, 2000; Pearl, 1998; Rosenbaum and

Rubin, 1983a). The main concern is about whether the distribution of missing observations of interest is independent of the assignment to the treatment. The propensity score is suggested as a solution to this problem by modeling all possible covariates to approximate missing values. However it is built on a strong assumption that all meaningful covariates in terms of theoretical and statistical significance should be exhausted in the model, which is quite a difficult condition to satisfy in practice. Even though it is done to quantify the probability of casual inference or the effect of selection bias, most proponents of this method admit that it is not easy to get the proper model specification from which propensity scores are obtained.

The difference between propensity scores and the index of impact of missing data proposed in this paper lies on whether or not auxiliary information for missing values is available. The index *M* is developed as a function of sample size, covariance and variances which are available from a given sample at hand. Further, the index is simply based on thresholds of statistical significance to evaluate the impact of missing data with a null relationship between X and Y rather than assuming a model for a predicted probability of being missing.

Nonparametric Model

Nonresponse or drop-out is a typical, difficult problem in longitudinal studies, which causes selection bias. Often, nonrespondents differ critically from respondents but the extent of that difference is unknown unless we can obtain follow-up information about the nonrespondents population (Lohr, 1999). So selection bias occurs when the sampled population is not the same as the target population.

A nonparametric (or semiparametric) model is to quantify the selection bias caused by

nonresponses (see, Daniels and Hogan, 2000; Robins, et. al., 2000; Scharfstein, et. al., 1999). The model is nonparametric in the sense that parameters indicating nonresponse mechanism are chosen rather than estimated from an obtained data. This implies that one can examine a wide range of approximates of missing data to evaluate the model specification. Selection bias parameters are usually defined by substantive area experts in that these parameters can not be identified directly from the distribution of the observed data. This model has been applied to longitudinal data sets in which each subject is measured at least once (e.g., studies in epidemeology).

While the nonparametric model starts from the fact that the missing mechanism is unknown it still bases the inference on the specified selection bias model that identifies the relation between response rates and covariates, and somehow designates selection parameters. Therefore it depends on whether or not the selection model is properly specified, and how reliable experts' hunch on the range of selection bias parameters is.

Compared with the approach of nonparametric models, neither any model with hypothetical parameters is assumed nor a sensitivity analysis is needed to develop and evaluate the index *M*. The index simply represents the impact of missing data relative to the cutpoint of statistical significance.

File Drawer Problems in Meta Analysis

In the realm of meta analysis, synthesists are concerned about the stability of statistical results from various primary studies which are reviewed (see, Begg, 1994; Brown, 1992; Thompson and Kieffer, 2000). So it has been asked "What if the synthesis had included more nonsignificant studies?" This concern is called publication bias: Journal editors are more likely

to accept papers that present significant statistical results while unpublished studies with nonsignificant results remain in researchers' file drawers. Therefore, synthesists more easily retrieve published studies than unpublished ones and then the synthesis of primary studies might be biased toward the statistical significance.

This concern about access to nonsignificant results is also called the file drawer problem and several statistical procedures have been developed to express the stability of the synthesis and seriousness of file drawer problem, fail-safe N (Orwin, 1983; Rosenthal, 1979). If the fail-safe N is very large compared to the number of primary studies included in the synthesis, a researcher is fairly assured the obtained results are robust.

Indeed, the file drawer problem pursues the same question that has been discussed in this paper, in that 1) the dominant statistic (e.g., effect sizes and covariance of supplemental data) is artificially fixed zero for unobserved cases, 2) the sample size (e.g., the number of primary studies and observations) necessary to alter the initial inference is calculated, and 3) both are based on the critical values of test statistics.

Compared with fail-safe N, the index M is a scale-free metric because it has a form of ratio of two sample sizes[11]. Also the index M could provide more tangible and situation-specific interpretation for a calculated number of unobserved/missing cases. While $5k+10$ (k is the number of primary studies in the synthesis) is suggested as a general and conservative tolerance level for the file drawer problem, we can evaluate the index M (or the number of unobserved cases) with the sample- or situation-specific information on response rates or population structures.

In sum, the ratio index M, is developed under the assumptions of constant means and

variances, and a null covariance of supplemental cases. These conditions make the index quantify the impact of the unobserved/missing data on the statistical inference. This index can be used as a benchmark when one wants to evaluate the stability of the initial statistical inference from the observed data. Analogous to the fail-safe N in meta analysis, this index, a ratio of sample sizes, may provide useful information about how likely the obtained statistical result from the observed data holds for the initially intended sample or across subgroups.

Although the proposed index can be used as a heuristic device especially for generalization of the sample and quantification of the impact of unobserved/missing cases, its sampling distribution is not specified. At this point, it is unknown yet whether this index is related to any known statistical distributions but a practical solution may be obtained from the empirical distribution or the reference distribution that is discussed by Frank (2000) in the context of confounding variables.

Notes

1. Little and Rubin (1989) also characterized missing patterns into several types (univariate missing, unit nonresponse, monotone missing and general pattern), which are again related to what statistical methods are appropriate to make better inferences.

2. The likelihood function consists of two parts, one for completely observed variables and the other for conditional distribution of incomplete variables.

3. For TIMSS, several groups of researcher and statisticians (International Study Center, Statistics Canada, the sampling referee, and the Technical Advisory Committee) evaluate the quality of the samples based on their own criteria (Dumais, 1998, p 15).

4. In another way, different response rates from different subgroups can be handled by adopting appropriate weighting methods.

5. Actually, values of the negative solution for M in equation 9 are very close to zero such that they are meaningless. For example, when n=84, negative solutions across all r's (-1 to 1) are -.05 to 0.

6. These two sample sizes correspond to a statistical power of .80 for the small (.10) and medium (.30) sizes of correlations at .05 level (Cohen and Cohen, 1983). Another sample size (n=783) corresponding to the large correlation (.50) is not included in Figure 3 for efficient graphic representation but it will be included in case 2.

7. In order to include the exact value of critical t, we may express the critical t values as a function of the combined sample size. Indeed, an inverse function of sample sizes has a high predictability for t values (more than 95% explanation). However, to use an inverse function of sample size makes equation 9 be a cubic function of M. For a third power equation, we can find exact solutions by using Cardan formulas. However, Cardan solutions are too complicated to get a handy index, so that they are not considered here.

8. The variable FAOCC is arbitrarily selected to present an example but not to provide any substantive argument.

9. Featherman and Hauser (1976) didn't provide specific information on the target and the sampled population.

10. When the interaction term is statistically significant, it means that the relationship between X and Y depends on subsamples and can not generalize across groups. Since we don't know values of X and Y for the unobserved, we can not includes an interaction term in the model, which indicates the difference between observed and unobserved cases. Instead, the index M might be a one way to evaluate the interaction effect in practice by assessing conditions which change the inference obtained from the observed data

11. However, Orwin's fail-safe N is easily expressed as a ratio of two sample sizes.

References

Allison, P. D. (2000). Multiple imputation for missing data. *Sociological Methods & Research*, 28 (3), 301-309.

Begg, C. B. (1994). Publication bias, In Harris Cooper and Larry V. Hedges (Eds.) *The Handbook of Research Synthesis*, NY: Russell Sage Foundation.

Birnbaum, M. H., and Mellers, B. A. (1989). Mediated models for the analysis of confounded variables and self-selected samples. *Journal of Educational Statistics*, 14 (2), 121-140.

Brown, J. R. (1992). Detecting potential hucksterism in meta-analysis using a follow-up fail safe test. *Psychology in the Schools*, 29, 179-184.

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Science*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. C., and Campbell, D. T. (1979). *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand Mcnally.

D'Agostino, R. B. Jr., and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95 (451), 749-759.

Daniels, M. J., and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56, 1241-1248.

Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94 (448), 1053-1062.

Dumais, J. (1998). Implementation of the TIMSS sampling design. In Martin, M. O. and Kelly, D. L. (Eds) *Third International Mathematics and Science Study, Technical Report*, Volume III: Implementation and Analysis, Final Year of Secondary School.

Featherman, D. L. and Hauser, R. M. (1976). Sexual inequalities and socioeconomic achievement in the U.S. 1962 – 1973. *American Sociological Review*, 41, 462-483.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29 (2), 147-194.

Frank, K. A. (2001). The Sensitivity of a Statistical Inference to Concerns about Unrepresented Populations. Unpublished manuscript.

Gourieroux, C., and Montfort, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies*, XLVIII, 579-586.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29, 722-729.

Guttman, I., and Menzefricke, U. (1983). Bayesian inference in multivariate regression with missing observations on the response variables. *Journal of Business and Economic Statistics*, 1, 239-248.

Hunter, J. E. (1997). Needed: A ban on the significance test. *American Psychological Society*, 8 (1), 128-135.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87 (420), 1227-1237.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing data*. New York: John Wiley.

Little, R. J. A., and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, 18 (2 & 3), 292-326.

Little, R. J. A., and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies

via potential outcomes: Concepts and analytical approach. *Annual Review Public Health*, 21, 121-145.

Lohr, S.L. (1999). Sampling: Design and Analysis. CA: Brooks/Cole.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal or Educational Statistics*, 8, 157-159.

Pearl, J. (1998). Graphs, causality, and structural equation model. *Sociological Methods & Research*, 27 (2), 226-284.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran and D. Berry (Eds) *The IMA Volumes in Mathematics and its Application*, 116, 1-96.

Rosenbaum, P. R., and Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effect. *Biometrica*, 70, 41-55.

Rosenbaum, P. R., and Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45 (2), 212-218.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of American Statistical Association*, 93 (444), *Theory and Method*.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of American Statistical Association*, 94 (448).

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implication for training of researchers. *Psychological Methods*, 1, 115-129.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55 (2), 455-466.

Thompson, B. (1989) Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counsleing and Development*, 22, 2-5.

Thompson, B., and Kieffer, K. M. (2000). Interpreting statistical significance test results: A proposed new "What if" method. *Research in the Schools*, 7(2), 3-10.

Wainer, H. (1989). Eelworms, bullet holes, and Geraldin Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics*, 14(2), 121-140.

BEST COPY AVAILABLE

Table 1. Correlation Coefficients among Education and Background Variables for Men in 1973 (Featherman and Hauser, 1976)

	Father's Occupation	Farm Origin	Number of Siblings	Education
Father's Occupation	1.00	-.412	-.289	.416
Farm Origin		1.00	.265	-.312
Number of Siblings			1.00	-.360
Education				1.00

Originally Observed Statistical Significance	Change of Sample: Yes	Change of Sample: No
Yes	Case 1 (Addition)	Case 2 (Replacement)
No	Case 3 (Removal)	Case 4 (Replacement)

Figure 1. Significance of $\hat{\beta}_1$, Change of Resultant Sample Sizes, and 4 Hypothetical Situations

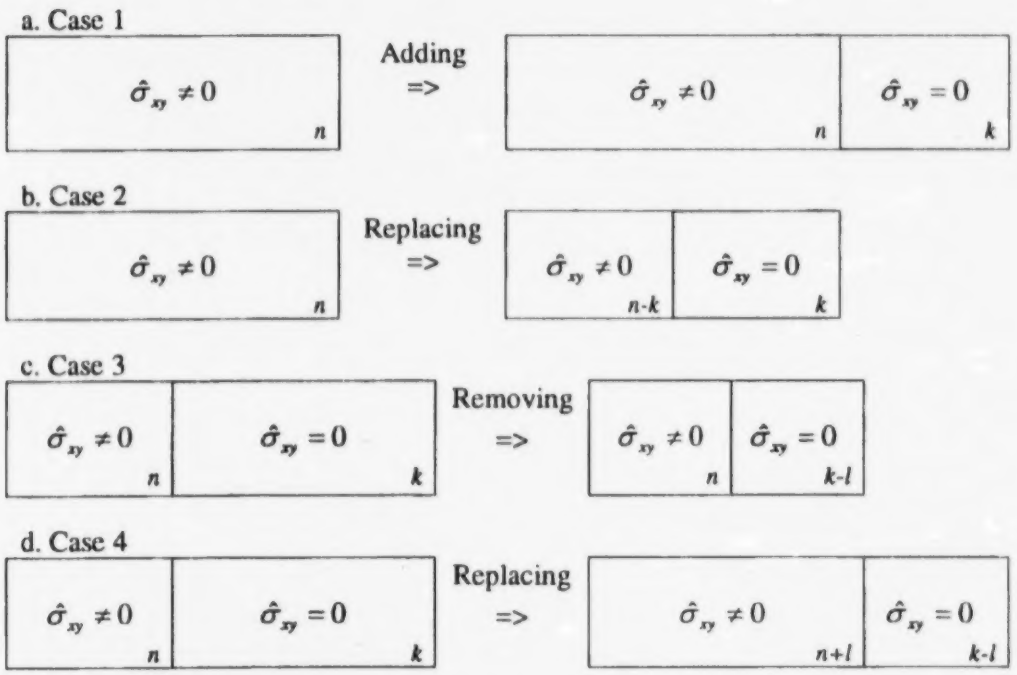

Figure 2. Data Structures for the Four Cases

BEST COPY AVAILABLE

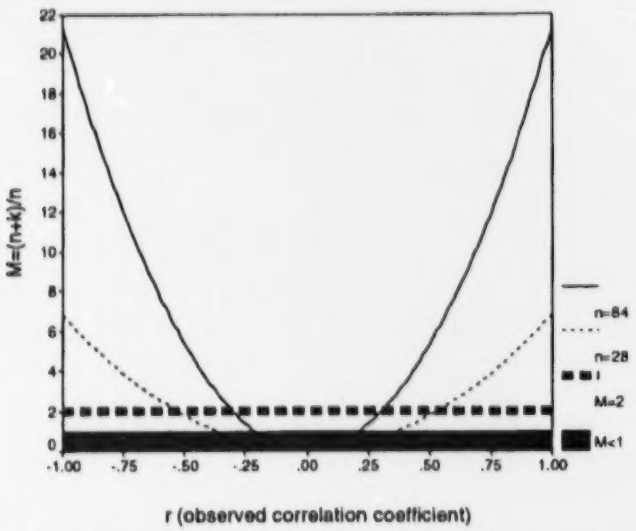

Figure 3. Index *M*, Correlation Coefficients (*r*), and Observed Sample Sizes (*n*) in Case 1 (α =.05)

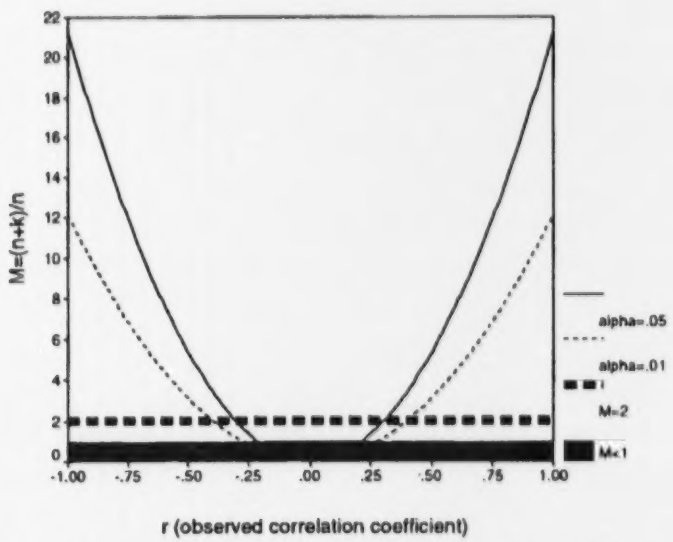

Figure 4. Index *M*, Correlation Coefficients (*r*), Significant Level α 's in Case 1 (*n*= 84)

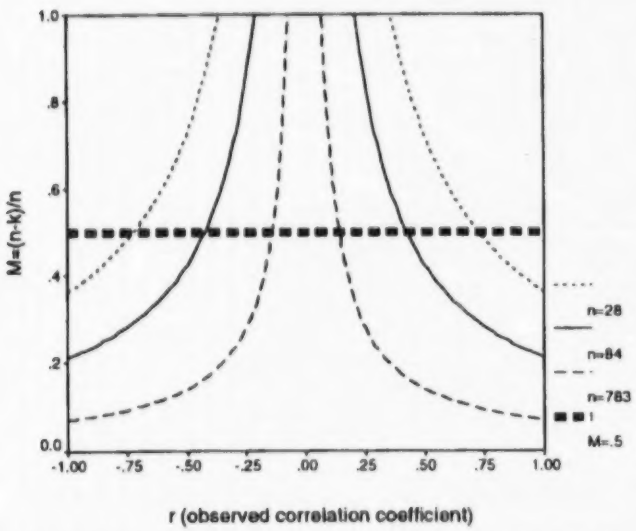

Figure 5. Index *M*, Correlation Coefficients (*r*), and Observed sample Sizes (*n*) in Case 2 (α =.05)

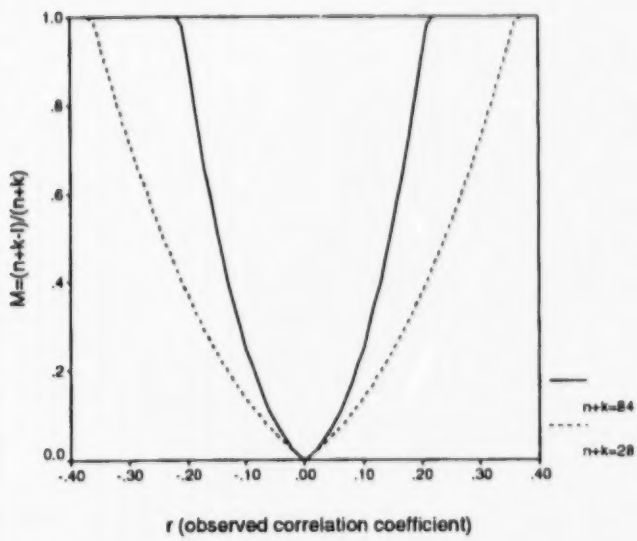

Figure 6. Index *M*, Correlation Coefficients (*r*), and Observed sample Sizes (*n*+*k*) in Case 3 (α =.05)

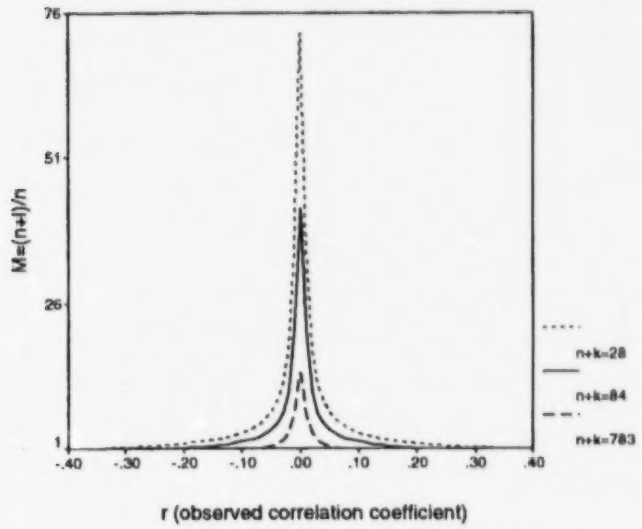

Figure 7. Index *M*, Correlation Coefficients (*r*), and Observed sample Sizes ($n+k$) in Case 4 (α =.05)*

* Because the index *M* can not be defined when the observed correlation is zero, .004 is used for zero correlation in this figure.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034785

I. DOCUMENT IDENTIFICATION:

Title: The Impact of Nonignorable Missing Data on the Inference of Regression Coefficients	
Author(s): Kyung-Seok Min and Kenneth A. Frank	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Min Kyung-Seok	Printed Name/Position/Title: Kyung-Seok Min,
Organization/Address: Michigan State University	Telephone: 517-432-2703 FAX:
East Lansing, MI 48824	E-Mail Address: minkyung@msu.edu Date: 1/30/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)